The measurement of neck pain and low back pain and the role of psychosocial factors in chiropractic care

Ailliet, L.

2016

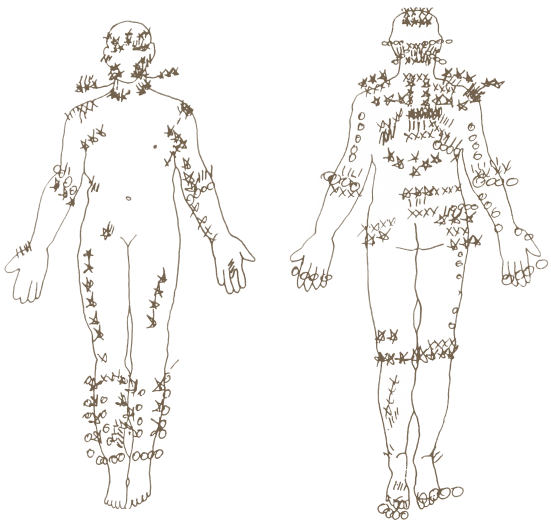
document version
Publisher's PDF, also known as Version of record

CHAPTER 4

Reliability, responsiveness and interpretability of the Neck Disability Index

Eur Spine J
(2015) 24:88–93

Ailliet L., Rubinstein S.M., van Tulder M., Terwee C.B., de Vet H.C.W.



ABSTRACT

Purpose

To establish an evidence-based recommendation for the pragmatic use of the Neck Disability Index-Dutch Version (NDI-DV) in primary care, based on an assessment of the reliability, the responsiveness, and the interpretability of the NDI-DV.

Study design and setting/methods

At baseline, the NDI-DV was completed by 337 patients with neck pain presenting to 97 chiropractic clinics in Belgium and the Netherlands. Three months after inclusion 265 patients provided data to assess the responsiveness and interpretability. Reliability was assessed in 155 patients (retested after 10 days) by calculating the intra-class correlation coefficient for agreement ($ICC_{agreement}$) and the measurement error (Standard Error of Measurement, SEM), the latter resulting in the smallest detectable change (SDC). The minimal important change (MIC) was assessed by the anchor-based MIC distribution using self-reported perceived recovery as anchor. We tested interpretability by relating SDC to MIC.

Results

The $ICC_{agreement}$ was 0.88. The $SEM_{agreement}$ was 1.95 resulting in a SDC of 5.40. The NDI-DV appeared to be responsive, being able to distinguish improved from stable patients with an area under the curve of 0.85. The MIC was 4.50.

Conclusion

The NDI-DV has good reliability and responsiveness and may be used in clinical practice in Belgium and The Netherlands. A change score of 5 is important for patients, but has a 7% chance to be due to measurement error.

Neck pain is a common musculoskeletal condition. Clinicians offer various therapies to patients who seek conservative care for neck pain. They assess the effect of a particular treatment by relevant outcome measures such as functional (dis)ability, pain and perceived recovery. Patient reported outcome measures (PROMs), defined as any report coming directly from patients about how they function or feel in relation to a health condition and its therapy, without the interpretation of the patient's responses by a clinician or anyone else[1], are commonly used for that purpose. The Neck Disability Index (NDI), the most frequently evaluated neck-specific questionnaire[2,3], is an example of a PROM.

As PROMs have become increasingly popular as measurement instruments both in clinical practice and epidemiological studies, there is a clear need to determine which scores or changes in scores on these questionnaires are important.

In this study we examine the domains reliability and responsiveness and the interpretability of the Dutch NDI (NDI-DV). Reliability, measurement error, responsiveness and interpretability are evaluated and interpreted according to the definitions set forth by the COnsensus-based Standards for the development of Measurement INstruments (COSMIN) panel[4].

To interpret change scores on a PROM, one needs to consider two bench marks: measurement error and minimal important change (MIC)[5]. Measurement error can be expressed as standard error of measurement (SEM) or as smallest detectable change (SDC). The MIC is the smallest change in score that patients consider important[6]. To explore the interpretability, the SDC is related to the MIC. To our knowledge, only one study in a tertiary care setting[7] has examined the measurement error (SDC) and the MIC of the Dutch version of the NDI. The goal of this study is to establish an evidence-based recommendation for the pragmatic use of the NDI-DV in primary care, based on an assessment of the reliability, the responsiveness, and the interpretability of the NDI-DV.

METHODS

Population

For the purpose of a prospective cohort study, patients with neck pain were recruited in 97 chiropractic clinics in Belgium and The Netherlands from August 26th to December 30th 2010. Inclusion criteria were: men and women, age 18 to 65, who had neck pain with or without radiation into the arm as their main complaint, had not consulted a chiropractor for their neck complaint in the past 6 months, had a good understanding of the Dutch language and had access to the internet.

Measurements

The questionnaires were sent electronically to the participating patients at baseline and 6 follow-up time points: at the second visit (on average 1 week after the baseline visit), after 1 month, 3 months, 6 months, 6 months + 10 days, and 12 months. The baseline data and the 3 months data were used to assess the responsiveness and interpretability, and the 6 months and 6 months + 10 days data were used to assess the reliability and measurement error.

The version that was used in previous studies on the NDI-DV was also used in this study[8]. The NDI consists of 10 items: pain intensity, personal care, lifting, reading, headache, concentration, work, driving, sleeping and recreation. The 10 items have 6 response categories (range 0-5, with 0 = no disability and 5 = total disability) resulting in a total sc ore range from 0-50 points with higher scores indicating more disability[9]. In addition, patients rated their current pain and average pain over the past week on a 0-10 numerical rating scale, and graded their perceived recovery since baseline. Perceived recovery was rated on a 7-point Likert Scale. We trichotomized this scale: patients that were somewhat better, not changed or somewhat worse were labeled as "not importantly changed". Patients who had indicated that they were much better or had completely recovered were labeled as "importantly improved" and patients who had indicated that they were much worse or had the worst imaginable pain were labeled as "importantly deteriorated".

Data analysis

RELIABILITY AND MEASUREMENT ERROR

The reliability of the NDI-DV was assessed by rating test-retest reliability and measurement error[10]. The test-retest interval was set at 10 days. As a parameter of reliability of the NDI-DV, the intra-class correlation coefficient ($ICC_{agreement}$) was computed using a two-way random effects model ($ICC_{agreement} = \sigma^2_p / [\sigma^2_p + \sigma^2_m + \sigma^2_r]$), where the error variance consists of a component representing the systematic difference between the two measurements (σ^2_m) and a component for the residual (random) error (σ^2_r)[11]. σ^2_p represents the differences between the "true" scores of patients. An ICC is expressed as a value between 0 and 1; a value > 0.70 is considered acceptable[12]. We quantified the measurement error by the Bland and Altman method and by calculating the standard error of measurement ($SEM_{agreement}$) from the ICC formula, by taking the square root of the error variance ($\sqrt{(\sigma^2_m + \sigma^2_r)}$). In the Bland and Altman plot, the mean difference between two measurements (here the NDI-DV values at 6 months and those at 6 months + 10 days) represents the systematic error and the standard deviation (SD) of this difference represents the random error. The 95% limits of agreement were defined as the mean difference between the measurements ± 1.96 x SD of the differences.

Smallest detectable change

The smallest detectable change (SDC) was based on the $SEM_{agreement}$. To be 95% confident that the observed change is not caused by measurement error but can be considered real change, the SDC at individual level (SDC_{ind}) was calculated as $1.96 \times \sqrt{2} \times SEM_{agreement}$. The SDC expresses the magnitude of change – with a probability of less than 5% – that this change is due to measurement error. Given this small probability, it is likely that a patient whose score exceeds the SDC has changed[13].

Responsiveness

The correlation between the anchor and the change scores on the NDI-DV was calculated. The area under the ROC curve (AUC) was computed. The AUC can be interpreted as the probability of correctly identifying an improved patient from randomly selected pairs of improved and stable patients[14]. A value > 0.70 for the AUC is considered satisfactory[6]. As responsiveness can be affected by the presence of floor and ceiling effects, the frequency of the highest and lowest possible scores at baseline and at the different follow-up measurements was assessed. Floor and ceiling effects can occur if more than 15% of the patients achieve the lowest or highest possible score at baselin[15]. For this purpose, we used the SDC, and defined the scale width in terms of not more than 15% of the respondents within 1 SDC value from the theoretical minimum or maximum of the scale[16].

Interpretability

The interpretability of the NDI for use in individual patients was tested by relating the SDC to the MIC. The SDC should be smaller than the MIC[13]. We determined the MIC by a ROC analysis using the 'perceived recovery'-scale as anchor. For the two groups, "importantly improved" and "not-importantly changed", the distribution of the change scores on the NDI are depicted in a graph, named the anchor-based MIC distribution[17].

The sensitivity values are plotted on the y-axis against the 1 – specificity values on the x-axis to distinguish patients who had improved from those who remained stable. To determine the MIC we defined the optimal cut-off point as the point that represents the lowest overall misclassification, i.e. where both sensitivity and 1 – specificity are maximized[17]. This ROC cut-off point was used to determine the proportion of "importantly improved" persons according to the anchor who are correctly identified by the NDI-DV as importantly improved (sensitivity) and to determine the proportion of "not-importantly changed" persons according to the anchor who are correctly identified by the NDI-DV as not-importantly changed (specificity).

All statistical analyses were performed in IBM SPSS statistics version 20.

RESULTS

Patient characteristics

At baseline, 337 patients completed the NDI-DV, 265 (78.6%) at the 3 months follow-up and 256 (76%) at the 6 months follow-up. Data from 265 patients (mean age 41.3 years, SD 11.8 years, 65.7% female) were used in the analysis to assess interpretability. Table 1 shows the baseline characteristics of the patients. At 3 months, 182 patients had importantly improved. The two groups – importantly improved and not importantly changed – were comparable in age and sex. The mean initial NDI-DV and pain scores, and the NDI-DV and pain scores at 3 months are presented in Table 2. The mean score and standard deviations of the NDI-DV at baseline and at 3 months for the 7 different categories of global perceived effect (GPE) and for the trichotomized categories are presented in Table 3.

Reliability and measurement error

At the measurement time point of 6 months + 10 days, 155 patients (60.5%) provided the information used to assess the reliability. An intra-class correlation coefficient (ICCagreement) of 0.88 was found.

The 95% limits of agreement, presented in a Bland-Altman graph (Figure 1), were between 5.60 and – 5.02. This means that by definition, 95% of the differences between repeated measurements lie between 5.60 and -5.02. The $SEM_{agreement}$ was 1.95. Based on this SEM, a SDC value of 5.40 was found (calculated as $1.96 \times \sqrt{2} \times SEM_{agreement}$).

TABLE 1: BASELINE CHARACTERISTICS OF THE NECK PAIN PATIENTS

Gender (male/female): (n = 337) (%)	34.3/65.7
Age – *Mean [SD]* – *Range*	 41.3 [SD 11.8] 18-65
Total scores NDI – *Mean [SD]* – *Range*	 12.89 [SD 6.17] 0-32
First episode of neck pain (yes)	17.7%
Duration of the complaint – *< 6 weeks* – *> 6 weeks* – *6 weeks – 3 months* – *> 3 months*	 25.4% 74.6% 15.9% 58.7%
Education – *No high school diploma* – *High school diploma* – *College/university degree* – *Post-university degree*	 25.1% 35.4% 35.6% 3.9%
Referral pattern – *MD / other health care person* – *Family / friends* – *Own initiative*	 25.8% 41.0% 33.2%

TABLE 2: CHARACTERISTICS OF THE PEOPLE THAT HAD IMPORTANTLY IMPROVED AT 3 MONTHS VERSUS THOSE WHO HAD NOT IMPORTANTLY CHANGED

n = 256	IMPORTANTLY IMPROVED	NOT IMPORTANTLY CHANGED
Gender (male/female) (%)	33.1/66.9	39/61
Age (mean [SD])	42.0 [11.3]	42.8 [12.0]
NDI-DV score at baseline (mean [SD]) – score from 0 to 50	12.4 [6.1]	12.6 [5.7]
NDI-DV score at 3 months (mean [SD]) – score from 0 to 50	5.3 [4.7]	13.1 [7.3]
Neck pain at baseline (mean [SD]) – score from 0 to 10	5.0 [2.0]	4.7 [2.3]
Neck pain at 3 months(mean [SD]) – score from 0 to 10	1.3 [1.7]	3.8 [2.3]

TABLE 3: MEAN SCORE AND STANDARD DEVIATIONS OF THE NDI AT BASELINE AND AT 3 MONTHS FOR CATEGORIES OF GLOBAL PERCEIVED RECOVERY (GPR AT 3 MONTHS)

CATEGORIES OF GPR	T0 (N = 337)	T3 (N = 256)	MEAN CHANGE
Completely recovered (*n* = 35)	11.4 [5.8]	1.1 [1.6]	10.3 [5.2]
Much better (*n* = 141)	12.8 [6.2]	6.4 [4.6]	6.4 [5.7]
Slightly better (*n* = 52)	12.5 [5.9]	13.0 [7.4]	-0.5 [4.1]
Unchanged (*n* = 13)	13.2 [6.4]	14.2 [8.0]	-1.0 [2.9]
Slightly worse (*n* = 7)	12.0 [3.8]	13.7 [4.8]	-1.7 [3.4]
Much worse (*n* = 0)	No data	No data	No data
The worst imaginable (*n* = 2)	15.0 [2.8]	22.0 [8.5]	-7.0 [5.7]
Total	12.9 [6.2]	7.9 [6.8]	5.0 [6.2]
Importantly improved (176)	12.5 [6.1]	5.3 [4.7]	7.2 [5.8]
Not importantly changed (72)	12.6 [5.7]	13.3 [7.3]	-0.7 [3.8]
Importantly worsened (*n* = 2)	15.0 [2.8]	22.0 [8.5]	-7.0 [5.7]

T0 baseline, *T3* at 3 months follow-up

Responsiveness

The correlation between the anchor (perceived recovery) and the change scores on the NDI-DV is 0.54. For responsiveness, the ROC analysis showed an area under the curve (AUC) of 0.85. The percentages of patients with NDI-DV scores between 0 and 5 at baseline and the different follow-up time points are presented in Table 4. By the time of the second visit – on average less than 1 week post the baseline measurement – more than 15% (17.5%) of the respondents scored ≤ 5/50. As the follow-up measurements progressed, the percentages of respondents scoring ≤ 5/50 increased, reaching 44.1% at the 6-months follow-up. Neither at baseline nor at the subsequent follow-up time points was the theoretical minimum score of the NDI-DV (0/50) scored by more than

FIGURE 1: BLAND AND ALTMAN PLOT ILLUSTRATING THE MEAN DIFFERENCE BETWEEN TWO MEASUREMENTS AND THE SD

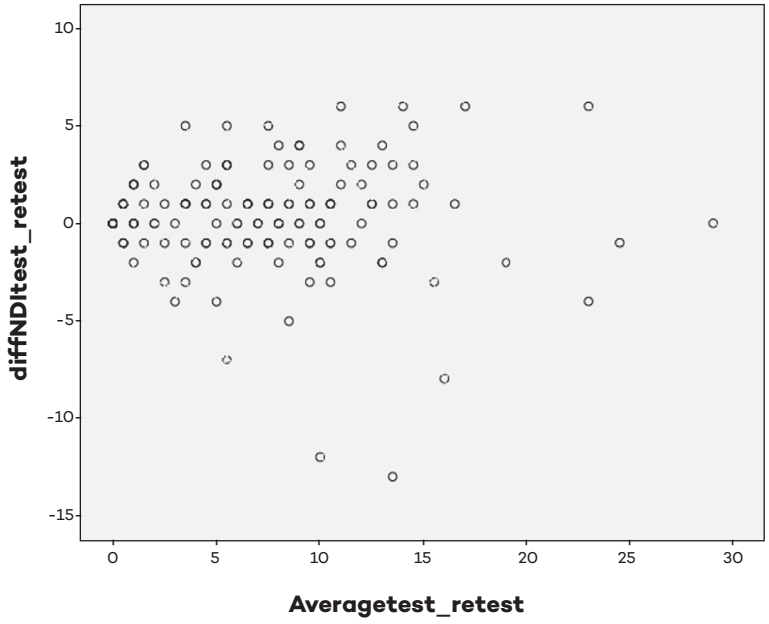


TABLE 4 ILLUSTRATING POTENTIAL FLOOR EFFECTS

NDI SCORE	T0 (%) (BASELINE)	2ND VISIT (%)	T1 (%) (1 MONTH)	T3 (%) (3 MONTHS)	T6 (%) (6 MONTHS)
0	0.9	2.9	5.9	11.7	13.7
5	9.5	17.5	36.3	43.4	44.1

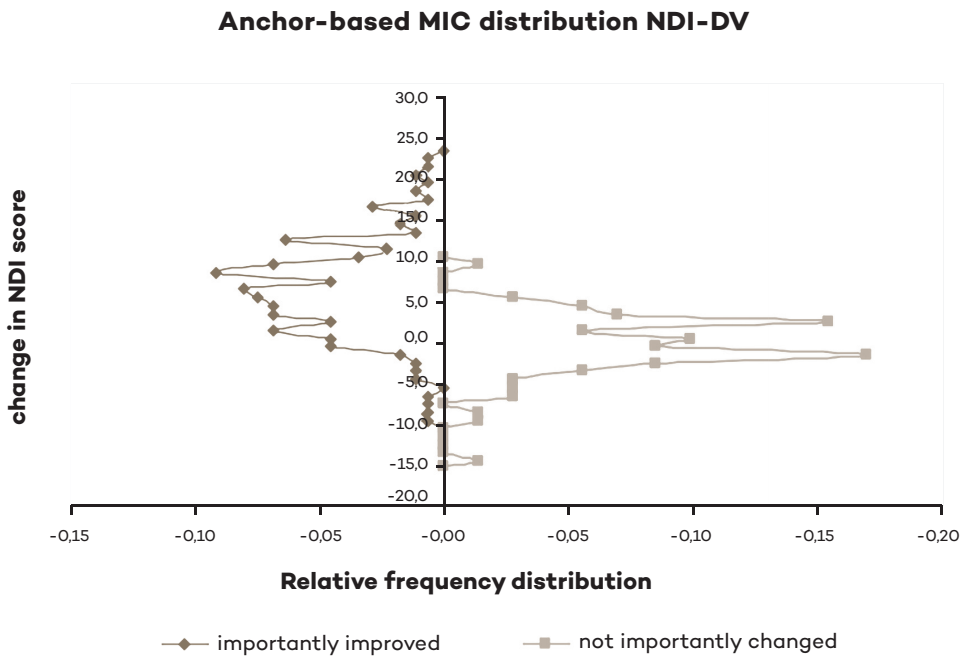
Percentages of patients with total NDI score 0 and total NDI score 5 at baseline, 2nd visit, 1, 3 and 6 months

15% of the respondents, with 13.7% at the 6-months follow-up measurement being the highest score. No patients, either at baseline or at any of the follow-up time points, presented with scores of 45/50 or 50/50.

Minimal important change

The optimal cut-off point (MIC value) corresponded to a score of 4.50. Figure 2 illustrates the anchor-based MIC distribution determining the MIC for the NDI-DV in patients with neck pain. With the MIC as cut-off point, 33 % of the anchor-based "importantly improved" patients had a lower change score and were considered false negatives. The

FIGURE 2: ANCHOR-BASED MIC DISTRIBUTION FOR THE NDI-DV



sensitivity of the NDI-DV was hence 0.67. Ten per cent of the anchor-based "not-importantly changed" patients had higher change scores than the cut-off point and were thus considered false positives. The specificity of NDI-DV was 0.90.

DISCUSSION

This study shows that the test-retest reliability of the NDI-DV, applied to a subgroup of patients presenting to a chiropractor with neck pain, is good. The ICC of 0.88 is consistent with previously reported values of 0.90[18] and 0.84[7] in primary care settings in the Netherlands.

We found evidence for good responsiveness of the NDI-DV (AUC = 0.85), in line with the findings of a previous study on the use of the NDI in primary care (0.83) in the Netherlands[18].

Since our study reports on both the SDC and the MIC, we can take a stand on the interpretability of the NDI-DV. In the past, divergent SDC values for the NDI-DV have

been reported, ranging from 7.62 in patients with acute neck pain (< 6 weeks)[18] to 10.4 in patients with non-specific neck pain (48% < 6 weeks and 26% > 13 weeks) in general practice[19]. The SDC value of 5.40 from our study agrees with the SDC value of 5.00 originally proposed by the author[11].

The MIC value was 4.50 in our study. Since scores on the NDI are expressed as whole numbers, this implies that a change score of 4 is not considered important by patients while a change score of 5 is. This value of 5 was slightly smaller than the SDC (5.40). Using a 90% confidence level instead of the 95% level, for an SEM of 1.95 the SDC would be 4.50 ($1.64 \times \sqrt{2} \times 1.95$, with 1.64 representing the standard normal deviate that corresponds to the 90% level) and thus equivalent to the MIC.

Failure to adequately address floor and ceiling effects in the NDI can result in suboptimal assessment of the functional status of many patients. As Table 4 indicates, as follow-up time progresses, increasing numbers of patients had a total NDI-DV score of 5 or less. There are now 2 possibilities: 1. If all patients scoring less than 5 indeed experience no or negligible neck disability, we do not say that there is a floor effect. 2. If they still have neck disability, but the NDI-DV does not pick this up, it is a shortcoming of the measurement instrument, and then we define this as "floor effect"[11]. Van der Velde[20] illustrated by means of a person-item threshold distribution graph that at the lower end of the scale there were sufficient items to discriminate between patients, meaning that the patients with a low score had no problems. As the patients in our study in terms of disability very much resemble the population in the Van der Velde study, we assume that the low scores after treatment were not due to a floor effect, but that the patients really had low scores. This is supported by the low scores on the pain outcome after three months.

Strengths and limitations: The results on reliability and responsiveness in our study are comparable to the results from other studies carried out in a primary care setting in The Netherlands. The analyses were carried out on a large sample, so that the numbers of cases who remained stable and those who improved were satisfactory. Since we used a web-based system, we were not confronted with missing values.

The anchor-based MIC distribution determining the MIC for the NDI-DV in patients with neck pain used "perceived recovery" as anchor. This anchor has been criticized for being a one-item question asking about change over a rather long time period[21]. For PROM's, the MIC should be considered from the perspective of the patient[11]. However, the decision about the MIC is often taken by the researcher and not by the patient, since it is the researcher who determines which category they define as minimally important. Although the correlation between the anchor and the change in scores on the NDI was only 0.54 – which is moderate – the change scores on the NDI were well able to distinguish patients who indicated to be importantly improved on the anchor, with an AUC of 0.85.

CONCLUSION

The reliability and responsiveness of the NDI-DV, applied to patients with non-specific neck pain in a chiropractic setting, are good. Considering a MIC value of 4.50 and SDC of 5.40, the NDI-DV could be used in clinical practice where a change score of 5 can be considered important for the patients with a less than 7% chance of being due to measurement error.

REFERENCES

1. Patrick DL, Burke LB, Powers JH (2007) Patient reported outcomes to support medical product labeling claims: FDA perspective. Value and Health 10(suppl2):125-137.
2. Schellingerhout JM, Verhagen AP, Heymans MW, Koes BW, de Vet HC, Terwee CB (2012) Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review. Qual Life Res 21(4):659-70.
3. Schellingerhout JM, Heymans MW, Verhagen AP, de Vet HC, Koes BW, Terwee CB (2011) Measurement properties of translated versions of neck-specific questionnaires: a systemic review. BMC Med Res Methodol Jun 6;11:87.
4. Mokkink LB, Terwee CB, Patrick DL et al. (2010) International consensus on taxonomy, terminology and definitions of measurement properties for health-related patient reported outcomes: result of the COSMIN study. J Clin Epidemiol 63:737-45.
5. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, Bouter LM, de Vet HCW (2007) Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 60:34-42.
6. de Vet HCW, Terwee CB, Mokkink LB, Knol DL (2011) Measurement in Medicine. Cambridge University Press, Cambridge UK.
7. Jorritsma W, de Vries GE, Dijkstra PU, Geertzen JHB, Reneman MF (2012) Detecting relevant changes and responsiveness of Neck Pain and Disability Scale and Neck Disability Index. Eur Spine J 21:2550–2557.
8. Köke AJA, Heuts PHTG, Vlayen JWS et al. (1996) Neck Disability Index. Pijn Kennis centrum Maastricht. Meetinstrumenten chronische pijn. Maastricht 52-54.
9. Vernon H, Mior S (1991) The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther 14:409-415.
10. De Vet HCW, Terwee CB, Knol DL, Bouter LM (2006) When to use agreement versus reliability measures. J Clin Epidemiol 59:1033-1039.

11. Macdermid JC, Walton DM, Avery S, Blanchard A, Etruw E, Mcalpine C, Goldsmith CH (2009) Measurement properties of the Neck Disability index: A Systematic Review. J Orthop Sports Phys Ther 39(5):400-417.
12. Nunally JC, Bernstein IH (1994) Psychometric theory. New York: McGraw-Hill Inc.
13. de Vet HCW, Ostelo RWJG, Terwee CB, van der Roer N, Knol DL, Beckerman H, Boers M, Bouter LM (2007) Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. Qual Life Res 16:131–142.
14. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29-36.
15. McHorney CA, Tarlov AR (1995) Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res 4:293-307.
16. Van der Roer N, Ostelo RW, Bekkering GE, Van Tulder MW, De Vet HCW (2006) Minimal clinically important change for pain intensity, functional status, and general health status in patients with nonspecific low back pain. Spine 31:578-582.
17. De Boer MR, De Vet HC, Terwee CB, Moll AC, Volker-Dieben HJ, Van Rens GH (2005) Changes to the subscales of two vision-related quality of life questionnaires are proposed. J Clin Epidemiol 58:1260-1268.
18. Vos CJ, Verhagen AP, Koes BW (2006) Reliability and responsiveness of the Dutch version of the Neck Disability Index in patients with acute neck pain in general practice. Eur Spine J 15:1729-1736.
19. Pool JJ, Ostelo RW, Hoving JL, Bouter LM, de Vet HCW (2007) Minimal clinically important change of the Neck Disability Index and the Numerical rating Scale for patients with neck pain. Spine 32:3047-3051.
20. van der Velde G, Beaton D, Hogg-Johnston S, Hurwitz E, Tennant A (2009) Rash analysis provides new insight into the measurement properties of the Neck Disability Index. Arthritis & Rheumatism 61(4):544-551.
21. Terwee CB, Roorda LD, Dekker J, Bierma-Zeinstra SM, Peat G, Jordan KP, Croft P, de Vet HCW (2010) Mind the MIC: large variation among populations and methods. J Clin Epidemiol 63:524-534.